

RECONFIGURABLE PROCESSOR MODULE COMPRISING HYBRID  
STACKED INTEGRATED CIRCUIT DIE ELEMENTS

RELATED APPLICATION

The present application is a Continuation-In-Part of U.S. Patent Application Serial No. 10/452,113 filed June 2, 2003, which is a Continuation of U.S. Patent Application Serial No. 10/012,057, now issued  
5 patent no. 6,627,985 filed December 5, 2001, both of which are incorporated herein by reference in their entirety and are assigned to the assignee of the present application.

10 BACKGROUND OF THE INVENTION

The present invention relates, in general, to the field of systems and methods for reconfigurable, or adaptive, data processing. More particularly, the present invention relates to an extremely compact  
15 reconfigurable processor module comprising hybrid stacked integrated circuit ("IC") die elements.

In addition to current commodity IC microprocessors, another type of processing element is commonly referred to as a reconfigurable, or adaptive,  
20 processor. These reconfigurable processors exhibit a number of advantages over commodity microprocessors in many applications. Rather than using the conventional "load/store" paradigm to execute an application using a set of limited functional resources as a  
25 microprocessor does, the reconfigurable processor actually creates the number of functional units it needs for each application in hardware. This results in greater parallelism and, thus, higher throughput for many applications. Conventionally, the ability for

a reconfigurable processor to alter its hardware compliment is typically accomplished through the use of some form of field programmable gate array ("FPGA") such as those produced by Altera Corporation, Xilinx, Inc., Lucent Technologies, Inc. and others.

In practice however, the application space over which such reconfigurable processors, (as well as hybrids combining both microprocessors and FPGAs) can be practically employed is limited by several factors.

Firstly, since FPGAs are less dense than microprocessors in terms of gate count, those packaged FPGAs having sufficient gates and pins to be employed as a general purpose reconfigurable processor ("GPRP"), are of necessity very large devices. This size factor alone may essentially prohibit their use in many portable applications.

Secondly, the time required to actually reconfigure the chips is on the order of many hundreds of milliseconds, and when used in conjunction with current microprocessor technologies, this amounts to a requirement of millions of processor clock cycles in order to complete the reconfiguration. As such, a high percentage of the GPRP's time is spent loading its configuration, which means the task it is performing must be relatively long-lived to maximize the time that it spends computing. This again limits its usefulness to applications that require the job not be context-switched. Context-switching is a process wherein the operating system will temporarily terminate a job that is currently running in order to process a job of higher priority. For the GPRP this would mean it would have to again reconfigure itself thereby wasting even more time.

Thirdly, since microprocessors derive much of their effective operational speed by operating on data in their cache, transferring a portion of a particular job to an attached GPRP would require moving data from the cache over the microprocessor's front side bus to the FPGA. Since this bus runs at about 25% of the cache bus speed, significant time is then consumed in moving data. This again effectively limits the reconfigurable processor to applications that have their data stored elsewhere in the system.

These three known limiting factors will only become increasingly significant as microprocessor speeds continue to increase. As a result, the throughput benefits that reconfigurable computing can offer to a hybrid system made up of existing, discrete microprocessors and FPGAs may be obviated or otherwise limited in its potential usefulness.

#### SUMMARY OF THE INVENTION

In accordance with the disclosure of a representative embodiment of the present invention, FPGAs, microprocessors and cache memory may be combined through the use of recently available wafer processing techniques to create a particularly advantageous form of hybrid, reconfigurable processor module that overcomes the limitations of present discrete, integrated circuit device implementations of GPRP systems. As disclosed herein, this new processor module may be conveniently denominated as a Stacked Die Hybrid ("SDH") Processor.

Tru-Si Technologies of Sunnyvale, CA (<http://www.trusi.com>) has developed a process wherein semiconductor wafers may be thinned to a point where metal contacts can traverse the thickness of the wafer

creating small bumps on the back side much like those of a BGA package. By using a technique of this type in the manufacture of microprocessor, cache memory and FPGA wafers, all three die, or combinations of two or  
5 more of them, may be advantageously assembled into a single very compact structure thus eliminating or ameliorating each of the enumerated known difficulties encountered with existing reconfigurable technology discussed above.

10 Moreover, since these differing die do not require wire bonding to interconnect, it is now also possible to place interconnect pads throughout the total area of the various die rather than just around their periphery. This then allows for many more  
15 connections between the die than could be achieved with any other known technique.

Particularly disclosed herein is a processor module with reconfigurable capability constructed by stacking and interconnecting bare die elements. In a  
20 particular embodiment disclosed herein, a processor module with reconfigurable capability may be constructed by stacking thinned die elements and interconnecting the same utilizing contacts that traverse the thickness of the die. As disclosed, such  
25 a processor module may comprise a microprocessor, memory and FPGA die stacked into a single block.

Also disclosed herein is a processor module with reconfigurable capability that may include, for example, a microprocessor, memory and FPGA die stacked  
30 into a single block for the purpose of accelerating the sharing of data between the microprocessor and FPGA. Such a processor module block configuration advantageously increases final assembly yield while concomitantly reducing final assembly cost.

Further disclosed herein is an FPGA module that uses stacking techniques to combine it with a memory die for the purpose of accelerating FPGA reconfiguration. In a particular embodiment disclosed herein, the FPGA module may employ stacking techniques to combine it with a memory die for the purpose of accelerating external memory references as well as to expand its on chip block memory.

Also further disclosed is an FPGA module that uses stacking techniques to combine it with other die for the purpose of providing test stimulus during manufacturing as well as expanding the FPGA's capacity and performance. The technique of the present invention may also be used to advantageously provide a memory or input/output ("I/O") module with reconfigurable capability that includes a memory or I/O controller and FPGA die stacked into a single block.

According to yet another embodiment of the invention, an alternative method eliminates any grinding or further mechanical steps and accomplishes the "stacking" of integrated circuit elements during the actual wafer fabrication process. As disclosed herein, this new processor module according to the alternative method of the present invention may be conveniently designated as a Stacked Integrated Circuit Function ("SICF") Processor.

Firstly, a base wafer is completely processed just as it would have been for any other conventional use to make, for example, a microprocessor. Contacts are distributed throughout the surface of the die. After the processing of the microprocessor die on the base wafer, the surface of the wafer is left in a

state that will allow for the further processing of other semiconductor elements having other functions.

Secondly, the base wafer is processed through all of the steps required to create an FPGA. The FPGA I/O metal layers connect to the surface contacts of the microprocessor just as if the die were physically thinned and stacked as previously described. The end result is that two die functions are stacked and interconnected as previously described, but done so by using wafer fabrication steps instead of mechanical discrete die thinning and stacking.

This alternative process according to a further embodiment of the invention allows for many different die functions to be stacked. In addition, functions that use different types of wafer processing steps can also be stacked since one die fabrication process is completed before the next process is started. Because wafer functions are completed before the next process is started, it is also possible to perform wafer level testing of one finished function before starting processing of the stacked function. This is beneficial for reducing test time of die after stacking subsequent functions since die failing and early stage could then be rejected. This also benefits in process optimization since failures are easily associated with a particular process step.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The aforementioned and other features and objects of the present invention and the manner of attaining them will become more apparent and the invention itself will be best understood by reference to the following description of a preferred embodiment taken

in conjunction with the accompanying drawings,  
wherein:

Fig. 1 is a simplified functional block diagram  
of a portion of a prior art computer system  
5 incorporating one or more multi-adaptive processing  
(MAP™ is a trademark of SRC Computers, Inc., Colorado  
Springs, CO) elements;

Fig. 2 is a more detailed, simplified functional  
block diagram of the multi-adaptive processing element  
10 illustrated in Fig. 1 illustrating the user logic  
block (which may comprise a field programmable gate  
array "FPGA") with its associated configuration read  
only memory ("ROM");

Fig. 3 is a functional block diagram of a  
15 representative configuration data bus comprising a  
number of static random access memory ("SRAM") cells  
distributed throughout the FPGA comprising the user  
logic block of Fig. 2;

Fig. 4 is a simplified, exploded isometric view  
20 of a reconfigurable processor module in accordance  
with the present invention comprising a hybrid device  
incorporating a number of stacked integrated circuit  
die elements;

Fig. 5 is a corresponding functional block  
25 diagram of the configuration cells of the  
reconfigurable processor module of Fig. 4 wherein the  
FPGA may be totally reconfigured in one clock cycle by  
updating all of the configuration cells in parallel;

Figs. 6-10 are sequential cross-sectional process  
30 flow diagrams in which two functional elements are  
fabricated on a single base wafer according to an  
alternative processing method of the present  
invention; and

Fig. 11 is a cross-sectional process flow diagram in which three functional elements are fabricated on a single base wafer according to an alternative processing method of the present invention.

5    DESCRIPTION OF A REPRESENTATIVE EMBODIMENT

With reference now to Fig. 1, a simplified functional block diagram of a portion of a prior art reconfigurable computer system 10 is shown. The computer system 10 incorporates, in pertinent part,  
10    one or more microprocessors 12, one or more multi-adaptive processing (MAP<sup>TM</sup>) elements 14 and an associated system memory 16. A system bus 18 bidirectionally couples a MAP element 14 to the microprocessor 12 by means of a bridge 22 as well as  
15    to the system memory 16 by means of a crossbar switch 24. Each MAP element 14 may also include one or more bidirectional connections 20 to other adjacent MAP elements 14 as shown

With reference additionally now to Fig. 2, a more  
20    detailed, simplified functional block diagram of the multi-adaptive processing element 14 illustrated in the preceding figure is shown. The multi-adaptive processing element 14 comprises, in pertinent part, a user logic block 32, which may comprise an FPGA  
25    together with its associated configuration ROM 34. A MAP control block 36 and associated direct memory access ("DMA") engine 38 as well as an on-board memory array 40 is coupled to the user logic block 32 as well as the system bus 18.

30    With reference additionally now to Fig. 3, a functional block diagram of a representative configuration data bus 50 is shown comprising a number of SRAM cells distributed throughout an FPGA



comprising the user logic block 32 of the preceding figure. In a conventional implementation, the configuration information that programs the functionality of the chip is held in SRAM cells distributed throughout the FPGA as shown.

Configuration data is loaded through a configuration data port 52 in a byte serial fashion and must configure the cells sequentially progressing through the entire array of logic cells 54 and associated configuration memory 56. It is the loading of this data through a relatively narrow, for example, 8 bit port that results in the long reconfiguration times.

With reference additionally now to Fig. 4, a simplified, exploded isometric view of a reconfigurable processor module 60 in accordance with a representative embodiment of the present invention is shown comprising a hybrid device incorporating a number of stacked integrated circuit die elements. In this particular implementation, the module 60 comprises a die package 62 to which is coupled a microprocessor die 64, memory die 66 and FPGA die 68, all of which have a number of corresponding contact points, or holes, 70 formed throughout the area of the package 62 and various die 64, 66 and 68. It should be noted that a module 60 in accordance with the present invention may also comprise any combination of one or more of the microprocessor die 64, memory die 66 or FPGA 68 with any other of a microprocessor die 64, memory die 66 or FPGA die 68.

During manufacture, the contact holes 70 are formed in the front side of the wafer and an insulating layer of oxide is added to separate the silicon from the metal. Upon completion of all front side processing, the wafer is thinned to expose the

through-silicon contacts. Using an atmospheric downstream plasma ("ADP") etching process developed by Tru-Si Technologies, the oxide is etched to expose the metal. Given that this etching process etches the  
5 silicon faster, the silicon remains insulated from the contacts.

By stacking die 64, 66 and 68 with through-silicon contacts as shown, the cache memory die 66 actually serves two purposes. The first of these is  
10 its traditional role of fast access memory. However in this new assembly it is accessible by both the microprocessor 64 and the FPGA 68 with equal speed. In those applications wherein the memory 66 is tri-ported, the bandwidth for the system can be further  
15 increased. This feature clearly solves a number of the problems inherent in existing reconfigurable computing systems and the capability of utilizing the memory die 66 for other functions is potentially very important.

With reference additionally now to Fig. 5, a  
20 corresponding functional block diagram of the configuration cells 80 of the reconfigurable processor module 60 of the preceding figure is shown wherein the FPGA 70 may be totally reconfigured in one clock cycle by updating all of the configuration cells in  
25 parallel. As opposed to the conventional implementation of Fig. 3, a wide configuration data port 82 is included to update the various logic cells 84 through an associated configuration memory 86 and buffer cell 88. The buffer cells 88 are preferably a  
30 portion of the memory die 66 (Fig. 4). In this manner, they can be loaded while the FPGA 68 comprising the logic cells 84 are in operation. This then enables the FPGA 68 to be totally reconfigured in one clock cycle with all of its configuration logic

cells 84 updated in parallel. Other methods for taking advantage of the significantly increased number of connections to the cache memory die 66 (Fig. 4) may include its use to totally replace the configuration bit storage on the FPGA die 68 as well as to provide  
5 larger block random access memory ("RAM") than can be offered within the FPGA die 68 itself.

In addition to these benefits, there is an added benefit of overall reduced power requirements and  
10 increased operational bandwidth. Because the various die 64, 66 and 68 (Fig. 4) have very short electrical paths between them, the signal levels can be reduced while at the same time the interconnect clock speeds can be increased.

15 Another feature of a system incorporating a reconfigurable processor module 60 is that the FPGA 68 can be configured in such a way as to provide test stimulus to the microprocessor 64, or other chips in the stack of the die package 62 during manufacture and  
20 prior to the completion of the module packaging. After test, the FPGA 68 can then be reconfigured for whatever function is desired. This then allows more thorough testing of the assembly earlier in the manufacturing process than could be otherwise achieved  
25 with traditional packaged part test systems thus reducing the costs of manufacturing.

It should be noted that although a single FPGA die 68 has been illustrated, two or more FPGA die 68 may be included in the reconfigurable module 60.  
30 Through the use of the through-die area array contacts 70, inter-cell connections currently limited to two dimensions of a single die, may be routed up and down the stack in three dimensions. This is not known to be possible with any other currently available stacking

techniques since they all require the stacking contacts to be located on the periphery of the die. In this fashion, the number of FPGA die 68 cells that may be accessed within a specified time period is

5 increased by up to  $4VT/3$ , where "V" is the propagation velocity of the wafer and "T" is the specified time of propagation.

Obviously these techniques are similarly applicable if other die types are added or substituted  
10 into the stack. These may include input/output ("I/O") application specific integrated circuits ("ASICs") or memory controllers and the like.

The disclosed technique for die interconnection used in forming the module of the present invention is  
15 superior to other available alternatives for several reasons. First, while it would be possible to stack pre-packaged components instead, the I/O connectivity between such parts would be much lower and limited to the parts' periphery, thereby obviating several of the  
20 advantages of the stacked die system disclosed.

Collocating multiple die on a planar substrate is another possible technique, but that too suffers from limited I/O connectivity and again does not allow for area connections between parts. Another option would  
25 be to fabricate a single die containing microprocessor, memory and FPGA. Such a die could use metalization layers to interconnect the three functions and achieve much of the benefits of die stacking. However such a die would be extremely large  
30 resulting in a much lower production yield than the three separate die used in a stacked configuration. In addition, stacking allows for a ready mix of technology families on different die as well as offering a mix of processor and FPGA numbers and

types. Attempting to effectuate this with a single large die would require differing mask sets for each combination, which would be very costly to implement.

According to an alternative method of the present invention, stacking of differing "functional elements" may be accomplished at wafer fabrication without any intervening grinding or other mechanical steps. While a particular method according to an embodiment of the present invention is shown and described, it will be apparent to those skilled in the art that many other variations of the described method are possible that will result in the stacking of different functional elements on top of and interconnected to each other as claimed.

As an example, it is desired to have a Field Programmable Gate Array (FPGA) function stacked on top of and interconnected to a microprocessor functional element. The first step in this sequence is to process a wafer in the normal fashion to create a microprocessor as shown in Fig. 6. The top surface of wafer 100 is a silicon dioxide surface 102 with metal pads 104 exposed to make electrical connections to the microprocessor circuits. The location of pads 104 can be anywhere on the surface of wafer 100. The wafer 100 is then ready for the second phase of the process.

In this phase, standard lithography and wafer processing techniques are used to mask off the metal contacts 104 and to grow an epitaxial layer of silicon on top of the existing oxide surface. A silicon dioxide layer 102 is shown in Fig. 7 that is formed after the metal contacts 104 are masked off. The epitaxial layer 106 grown on the silicon dioxide surface is shown in Fig. 8. Since epitaxial layer 106 is being grown on oxide layer 102, the resulting

epitaxial layer 106 is polycrystalline in nature. However, it is now possible to fabricate an FPGA function using polysilicon transistors in this epitaxial layer 106 using standard polysilicon processing techniques. During this process, the metal interconnects 108 of the FPGA are allowed to make electrical connection to at least one of the metal pads of the original microprocessor top surface as shown in Fig. 9. This can be done by etching vias through the polysilicon epitaxial layer 106 and underlying silicon dioxide layer 102 where an interconnection between functional elements is required, and forming the interconnect in the etched vias. In this fashion the microprocessor and the FPGA are physically coupled to create the desired stacked processor according to an embodiment of the present invention. When all of the FPGA process steps are complete, the top surface will again result in exposed metal pads 108 on a silicon oxide surface 110 as shown in Fig. 10. These pads 108 can then be used to interconnect to package bonding wires or flip chip solder balls using standard techniques for creating packaged integrated circuits.

Alternatively, pads 108 can be masked off and a further epitaxial layer can be formed so that a third functional element such as an I/O controller, memory, FPGA, microprocessor or the like, can be integrated with the first two functional elements. Referring now to Fig. 11, a module is shown wherein three functional elements are fabricated on a single base wafer, including functional elements 100, 106, and 112, separated by oxide layers 102 and 110. A surface oxide layer 114 is also shown. Metal pads 108 are shown interconnecting all three functional elements.

While there have been described above the principles of the present invention in conjunction with specific integrated circuit die elements and configurations for a specific application, it is to be  
5 clearly understood that the foregoing description is made only by way of example and not as a limitation to the scope of the invention. Particularly, it is recognized that the teachings of the foregoing disclosure will suggest other modifications to those  
10 persons skilled in the relevant art. Such modifications may involve other features which are already known per se and which may be used instead of or in addition to features already described herein. Although claims have been formulated in this  
15 application to particular combinations of features, it should be understood that the scope of the disclosure herein also includes any novel feature or any novel combination of features disclosed either explicitly or implicitly or any generalization or modification  
20 thereof which would be apparent to persons skilled in the relevant art, whether or not such relates to the same invention as presently claimed in any claim and whether or not it mitigates any or all of the same technical problems as confronted by the present  
25 invention. The applicants hereby reserve the right to formulate new claims to such features and/or combinations of such features during the prosecution of the present application or of any further application derived therefrom.

30 What is claimed is: